Final Report: Research Objects

Developing a workflow for ingest of Research Objects into DataBank

Jamie Wittenberg

Graduate School of Library and Information Science at the University of Illinois at Urbana-Champaign
e-Research Centre and Bodleian Library at Oxford University

Prepared October 2014
Updated January 2015

1. Context

Data-intensive scientific research is essential to discovery and advancement in nearly every domain. However, traditional models of scientific scholarly communication often limit researchers to publishing only aspects of studies that can be captured in text and static images. These models are inadequate for supporting reproducibility, a core tenet of the scientific process. Without instruments of the original research, experiments may be difficult or impossible to replicate, verify, and evaluate (Belhajjame et. al. 2014, p. 2). There is a need for an environment within which scientists can manage and exchange scholarly outputs that do not conform to traditional standards of research publication, but are crucial for interpreting, verifying, and reviewing results.

myExperiment.org, A joint project from the universities of Southampton, Manchester and Oxford that launched in 2007, addresses this need by providing an environment that enables the publication of in silico experiments and facilitates the aggregation of scientific research output. It allows researchers to bundle files, datasets, and executable workflows into mutable "Packs" that can be shared, exchanged, reused, and remixed. The impetus for establishing the internship project culminating in the preparation of this report stems from the need for archiving and referencing Packs in scholarly communication.

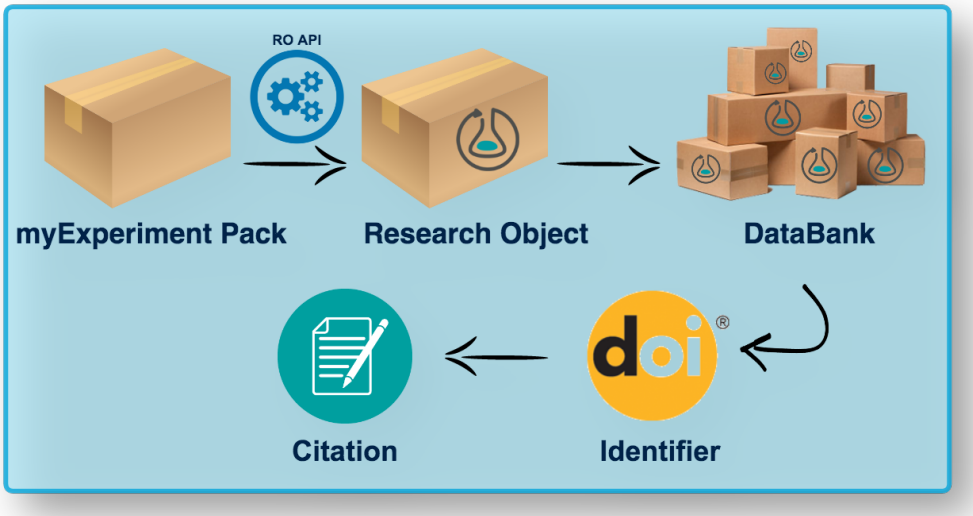

Fig. 1. Stages of the pack to Research Object lifecycle.

2. Proposed Workflow

The proposed lifecycle begins with a user-initiated deposit of a Pack from the myExperiment public-facing site. There are three stages (described below) in the manufacture of a citable archived pack (Fig. 1). A myexperiment.org pack is

archived by the user, triggering the creation of a Research Object (RO) using an RO API, which is processed for deposit. The RO is then ingested into DataBank where its metadata is indexed for discovery and a DOI is issued at the object level. Each stage is articulated in detail in sections 2.1 – 2.3.

2.1. Mapping a pack to an RO

The first stage in a Pack metamorphosis is to formalize its structure by mapping it to an RO. In their 2011 paper "Scientific social objects: The social objects and multidimensional network of the myExperiment website," De Roure et. al. explicitly address the need to generalize a myExperiment Pack to the notion of an RO. ROs can be approached as a set of principles related to identity, aggregation, and annotation. The Research Object Model, published by the Workflow Forever project (Wf4Ever), describes ROs as "semantically rich aggregations of resources that bring together data, methods and people in scientific investigations. Their goal is to create a class of artifacts that can encapsulate our digital knowledge and provide a mechanism for sharing and discovering assets of reusable research and scientific knowledge." The crucial elements of an RO in the context of archives ingest processes are identity, aggregation, and annotation. The object structure must knit together component parts, ensure identifiability and fixity at the object level, and expose the annotations embedded in the object.

2.1.1. Generating an RO

It is possible to generate an RO from a pack using existing tools and infrastructure. myExperiment Alpha 2 already implements an RO API developed by the Wf4Ever project to generate ROs from myExperiment packs (Fig. 2). Alpha 2 is still in development. The Research Object Digital Library (www.rohub.org) has deployed a more complete implementation of the RO API and samples were taken from that library to test DataBank ingest of ROs. Therefore, it is possible that there may be slight differences for myExperiment ROs.

Fig 2. Screenshot of exisiting RO generation via myExperiment Alpha2.

2.2. *Depositing an RO into DataBank*

Oxford University's Bodleian Library launched DataBank version 1.0.2, a repository for digital research data, in May 2013 . Bechhofer, Buchan, De Rour et al. describe three states in the lifecycle of a Research Object: Live Objects, Publication Objects, and Archived Objects. Live objects are mutable works that are still in development, publication objects represent distinct, citable works, and archived objects are immutable final products. Only objects at the end of their lifecycle in an Archived state are suitable for deposit into DataBank. DataBank has an API service that will allow front-ends such as ORA-Data to retrieve and ingest data into DataBank. Relying on existing data storage infrastructure reduces the cost of archiving Packs, and offers a trusted site of retrieval for the end user.

2.2.1. *DataBank Interoperability*

DataBank does support some nesting of files, including compressed files, in its data packages. It allows submitters to upload their own metadata in RDF/XML by including an RDF manifest file that merges with the DataBank package-level metadata created for each new data package. ROs also have an RDF manifest file that contains metadata on resources and annotations aggregated in the RO. Ensuring that DataBank is enabled to expose RO manifests protects the integrity of the RO's annotation layer, which is critical to understanding and using the object. One of the outcomes of this project was the opportunity to communicate this need to the DataBank development team.

2.3. *Assigning Unique Identification*

Widespread usage of the Digital Object Identifier System standard (ISO 26324:2012) coupled with the Bodleian Library's role as a DataCite DOI registration agency make DOIs an advantageous choice of identifier scheme. DOIs are infinitely scalable, supported by DataCite, have minimal implementation costs, and have been adopted by many commercial publishers (Duerr et. al. 2011). In the case that DOIs already exist for items bundled into a myExperiment pack, those identifiers should be documented in the RO manifest as related works either by the researcher at pack creation or during pre-processing, DOIs will resolve to DataBank landing page for an RO.

3. Recommendations

3.1. *Finalize myExperiement Alpha2*

Debugging and launching the myExperiment Alpha2 Implementation of the RO API would integrate packs and ROs. When a user initiates the deposit of a pack in an archived state- no longer a live, mutable object- the RO API should generate an RO that is processed before deposit into databank where the manifest is indexed for discovery and a DOI is issued.

3.2. *Create a Customized Search Index*

After consultation with Bodleian Digital Library Systems and Services, it was concluded that merging valid RDF RO manifests with Databank manifests should not be prioritized because the manifests can remain separate and continue to serve their intended functions. The RO manifest can be use to create a customized search index for ROs in DataBank. Creating a customized search index from DataBank manifests will allow for greater precision and improved functionality for end users retrieving ROs. A customized index could exploit the RDF structure of RO metadata and allow users to search based on types of relationships within and across ROs.

3.3. *Implement Pre-Processing of ROs*

Pre-processing ROs would allow for zip files containing other bundled objects to be unpacked and the RO manifest to be updated to integrate any missing metadata, including additional structural metadata and unique identifiers within bundled files. Pre-processing could also create a pre-deposit staging area that allows content creators to manipulate metadata and annotations or perform quality control prior to ingest. Pre-processing could be automated in situations where all normalization is consistent, or could be conducted on a case-by-case basis, allowing researchers to make curatorial decisions regarding the deposit. Pre-processing would take place immediately following RO generation and prior to ingest.

4. Future Work

4.1. *External objects*

External items associated with a pack do not travel with the pack when it is downloaded or transformed into an RO. These items are referred to as external objects, and though they are itemized in the manifest, they are not included in the bundle. Occasionally, external objects point to other packs in an archived state. Working to resolve this could have copyright and attribution issues, but it is crucial to approach those (potentially mutable) objects as being integral to the Pack. If external objects are not archived, this could be a bottleneck for verification and reproduction of research in the future.

4.2. *Provenance*

The relationships between entities captured in a myExperiemnt pack are object-centered. Agent-centered and process-centered provenance data are not systematically incorporated into the resultant research objects. Though event based metadata is less useful while researchers manipulate live packs in myExperiment, within the context of an archive, provenance data is critical to the trustworthiness of the record. An area for future research may be capturing and embedding provenance metadata in archived RO. The W3C PROV model may be a starting point for this work (Fig. 3).

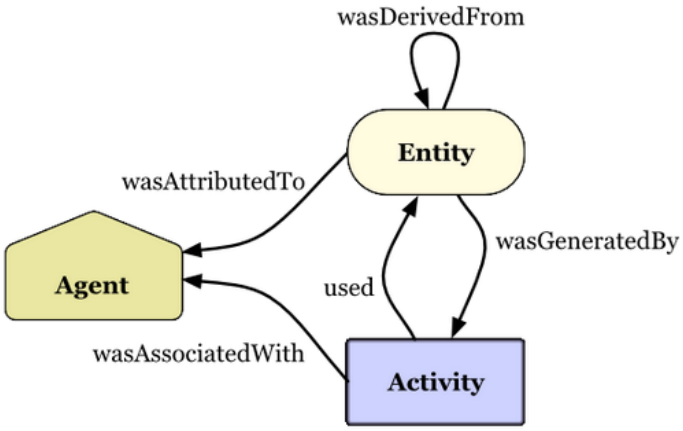

Fig. 3. W3C PROV Model.

5. Summary of Work Completed

To develop my proposed pack workflow and set of recommendations for RO submission information packages, I interviewed stakeholders including the Oxford e-Research Centre myExperiment team, Bodleian Digital Library Systems and Services, and Research Object model developer Graham Klyne. I assessed options for RO deposit workflows with a focus on the nuances of manipulating the existing databank web interface and merging RO manifests into DataBank manifests to expose structural metadata. The project deliverables include the final report and final presentation.

6. Acknowledgements

Generous financial support and project guidance were provided by the Center for Informatics Research in Science and Scholarship and the Graduate School of Library and Information Science at the University of Illinois as well as the Bodleian Libraries and the Oxford e-Research Center at Oxford University. I would like to gratefully acknowledge the supervision and assistance of Dr. David De Roure, Megan Senseney, Kevin Page, John Pybus, and Ruth Kirkham. Furthermore, I would like to acknowledge valuable input from Graham Klyne, Neil Jefferies, and Amanda Flynn.

7. References

Belhajjame, K., Zhao, J., Garijo, D., Hettne, K. M., Palma, R., Corcho, Ó… Goble, C.A. (2014). The Research Object Suite of Ontologies: Sharing and Exchanging Research Data and Methods on the Open Web. CoRR abs/1401.4307

De Roure, D. Bechhofer, S., Goble, C. A., & Newman, D. R. (2011). Scientific social objects: The social objects and multidimensional network of the myExperiment

website. *Privacy, Security, Risk and Trust (PASSAT), IEEE Third International Conference on Social Computing (SocialCom).* Boston, MA.

Duerr, R. E., Downs, R. R., Tilmes, C., Barkstrom, B., Lenhardt, W. C., Glassy, J., Bermudez, L. E., et al. (2011). On the utility of identification schemes for digital earth science data: an assessment and recommendations. Earth Science Informatics, 1–22.

wf4ever. (n.d.). *Research Object Model*. Retrieved October 13, 2014, from http://www.wf4ever-project.org/research-object-model